Word Distribution Analysis for Relevance Ranking and Query Expansion

Patricio Galeas and Bernd Freisleben*

Dept. of Mathematics and Computer Science, University of Marburg,
Hans-Meerwein-Str. 3, D-35032 Marburg, Germany
{galeas,freisleb}@informatik.uni-marburg.de

Abstract. Apart from the frequency of terms in a document collection, the distribution of words plays an important role in determining the relevance of documents for a given search query. In this paper, *word distribution analysis* as a novel approach for using descriptive statistics to calculate a compressed representation of word positions in a document corpus is introduced. Based on this statistical approximation, two methods for improving the evaluation of document relevance are proposed: (a) a relevance ranking procedure based on how query terms are distributed over initially retrieved documents, and (b) a query expansion technique based on overlapping the distributions of terms in the top-ranked documents. Experimental results obtained for the TREC-8 document collection demonstrate that the proposed approach leads to an improvement of about 6.6% over the term frequency/inverse document frequency weighting scheme without applying query reformulation or relevance feedback techniques.

1 Introduction

In a typical information search process, results are obtained by literally matching terms in documents with those of a query. However, due to *synonymy* and *polysemy,* lexical matching methods are likely to be inaccurate when they are used to meet a user's information need [1].

One way to address this problem is to consider contextual information [2]. In fact, several search engines make use of contextual information to disambiguate query terms [3]. Contextual information is either derived from the user, the document structure or from the text itself by performing some form of statistical analysis, such as counting the frequency and/of distance of words.

In this paper, we present an information retrieval approach that incorporates novel contextual analysis and document ranking methods. The proposed approach, called *word distribution analysis*, is based on a compressed statistical description of the word positions in a document collection, represented through their measures of *center* and *spread*. As a complement to the term frequency/inverse document frequency (*tfidf*) metric, we propose the *term density*

* This work is partially supported by the Deutsche Forschungsgemeinschaft (DFG, SFB/FK 615)